J. Math. Biol. 49: 455–467 (2004)
Digital Object Identifier (DOI):
10.1007/s00285-004-0271-9

Mathematical Biology

Stephan Morgenthaler · Pablo Herrero · William G. Thilly

Multistage carcinogenesis and the fraction at risk

Received: 3 March 2003 / Revised version: 19 November 2003 /
Published online: 23 April 2004 – © Springer-Verlag 2004

Abstract. Multistage carcinogenesis models describe the evolution of the cells in an individual's organ from a normal stage to a pre-neoplastic stage to a neoplastic stage. The triggers for the passage from one stage to the next one are presumed to be genetic alterations, which are not only governed by purely random events but also by individual environmental and genetic factors. We generalize existing models of carcinogenesis to populations composed of heterogeneous individuals, thus taking the environmental and genetic variability into account.

1. Introduction

Carcinogenesis can be modeled by an evolutionary process in which the cells in an organ divide and die (cell turnover) and genetic changes occur seemingly randomly. Specific genetic changes can produce increased cellular growth that might ultimately lead to neoplasia in the affected cells. If, for example, the inactivation of a gene is necessary for reaching a pre-neoplastic state, two mutations would be required to initiate the carcinogenesis in individuals who are homozygous in that gene. After reaching the pre-neoplastic stage, further genetic alterations in the cell may be necessary. To specify such a model we must make assumptions about the number of steps necessary for reaching the neoplastic state, rates of genetic change per cell division and information about the number of cells and their rates of division. Pioneering work in the area of carcinogenesis includes Nordling (1953) and Armitage and Doll (1954) and led to the development of the models described in Moolgavkar and Venzon (1979) and Moolgavkar and Knudson (1981), which have found wide acceptance. Further references and a collection of papers related to the topic discussed in this manuscript can be found in Moolgavkar (1990). A two-stage carcinogenesis model can be described by a series of linked stochastic processes, one for the normal cells, a second one for the initiated pre-neoplastic cells and a third one for the promoted neoplastic cells. Tan (1991) gives a detailed account of the mathematical aspects of these models.

S. Morgenthaler: Swiss Federal Institute of Technology (EPFL), SB/IMA, 1015 Lausanne, Switzerland. e-mail: stephan.morgenthaler@epfl.ch

P. Herrero, W.G. Thilly: M.I.T., Center for Environmental Health Sciences, Bldg. 16, Cambridge, MA 02139, USA

Key words and phrases: Competing risks – Age-dependent incidence rates – Turn over of cancer incidences at high age

We are interested in applying these models to gather insights into the process of cancer development in humans. Data on cancers, collected by cancer registries all over the world, are most often in the form of mortality rates due to specific cancers within geographic regions and well-defined populations. Carcinogenesis on the other hand describes the evolution of the disease in the organ of some individual. To link the available cancer data to a carcinogenesis model, we use the age-dependent incidence rate. To estimate this quantity in a population, we take a ratio

$$\text{obs}(t) = \frac{\text{number of incidences among persons of age } t}{\text{number of persons of age } t},$$

that is the fraction of incidences among the persons alive at age t. In order to get stable estimates, ages must be binned and we will usually work with the ages grouped into 5 year intervals, classes 0-5, 5-10, 10-15, 15-20, and so on. When working with the model on the other hand, it is more natural to start by computing the survival function $S(t)$, which is defined as the probability that the random process evolving in a randomly chosen individual has not yet produced a neoplastic cell at age t. The age-specific incidence rate is then equal to the hazard rate and satisfies

$$\text{pobs}(t) = -\frac{S'(t)}{S(t)}, \tag{1}$$

and inversely

$$S(t) = \exp\left(-\int_0^t \text{pobs}(u)\,du\right).$$

To use a model, we also need to make assumptions about the statistical variation of the model parameters from one individual in the population to another. Further complications in comparing data with a model could result from the difference between incidence and mortality, mistakes in death certificates, competing risks and similar uncertainties, but these problems will not be discussed further in this paper.

In this paper, we formulate a two-stage model for cancer incidence using the minimal requirements that our current understanding of carcinogenesis demand. In Section 2 we present this model and the relevant formulae for the implied incidence rates. In Section 3, we incorporate the fraction at risk into our model and illustrate how this parameter acts on the incidence rates.

2. Two-stage carcinogenesis

In a two-stage model, the normal cells of a tissue undergo an initiation that produces an irreversible change and leads to a growth advantage. Once an initiated cell is created, it gives rise to a cluster of such cells, which by random fluctuation may die or, again by chance, may undergo a promotion process which induces the further changes necessary for transforming them into neoplastic cells. The reader who is interested in obtaining more information about the mathematical modeling tools is referred to Todorovic (1992) and Kimmel and Axelrod (2002).

2.1. Initiation

For modeling purposes, it is not essential to know at which loci the mutations happen. Only the rate per cell division of each mutation and, of course, the number of necessary mutations are needed. The mutation rate per cell division is defined as the probability that after division one of the daughter cells carries the mutation. We work with the quantities $N(t)$, the number of normal cells at age t and $I(t)$, the number of initiated cells at age t. At birth, $N(0) = N_0$ and, unless a genetic defect (germ-line mutation) is present, $I(0) = 0$. If the number of necessary mutations is equal to one and on average each cell divides τ times per year and the mutation rate per cell division is r_1, the expected value of $I(t)$ is equal to

$$E(I(t)) = \int_0^t \tau r_1 N(u)\, du\,,$$

where we treat the number of cells $N(t)$ as a deterministic non-random function as in Coldman and Goldie (1983). The number of initiated cells $I(t)$ on the other hand is small and follows approximately a Poisson distribution with expectation $E(I(t))$. In considering the random trajectory of $I(t)$ as a function of t, we obtain a step function, that steps upward whenever a newly mutated cell appears. We can model this process as an inhomogeneous Poisson process with intensity

$$\lambda_I(t) = \frac{d}{dt} E(I(t)) = \tau r_1 N(t)\,.$$

If two mutations are required, the second one following the first one and ocurring at rate r_2, we have

$$E(I(t)) = \int_0^t \tau r_2 \left(\int_0^u \tau r_1 N(v)\, dv \right) du\,.$$

If the order in which the mutations occur does not matter, this value must be doubled and the corresponding rate of the Poisson process is

$$\lambda_I(t) = 2\tau^2 r_1 r_2 \int_0^t N(v)\, dv\,.$$

Each additional mutation adds a further integration. Under the assumption that $N(t) = N_0$ is constant over time, we find for n initiating mutations that can occur in any order

$$\lambda_I(t) = n\, \tau^n\, (r_1 r_2 \cdots r_n)\, N_0\, t^{n-1} \tag{2}$$

An mathematically more elegant model, based on a time-continuous branching process $N(t)$ is discussed in Section 4.2 of Kimmel and Axelrod (2002). The effects of this change on the final results are, however, negligable.

Once a cell is initiated, it remains so. This is an important implicit assumption we used in the above development. In organs, this would only hold for stem cells. To be realistic, one thus has to assume that the first $n - 1$ mutations happen in a stem cell, where they can be preserved. The last mutation may happen in any of the descendants of the stem cell in question. This has no fundamental importance in our formula.

2.2. Promotion

We first study what happens to an initiated cell created at some arbitrary age i. Such a cell has increased growth and thus will give rise to a clonal expansion. In its simplest form this can be modeled by a birth-and-death process having birth rate β greater than death rate δ. Let $C(t-i)$ be the (random) number of cells in the clone at time $t \geq i$. To keep things simple, we assume that each cell in the clone divides with the same birth rate β and disappears with the same death rate δ and acts independently of the other cells. In the birth case, an additional cell appears, that is a transition from $C(t-i)$ to $C(t-i)+1$ occurs, whereas in the death case $C(t-i)$ changes to $C(t-i)-1$. The probability for the first type of transition is

$$P\{C(t-i+h) = C(t-i)+1 \,|\, C(t-i)=k\} = k\,\beta\,h + o(h)\,.$$

The analogous equation for the other transition is

$$P\{C(t-i+h) = C(t-i)-1 \,|\, C(t-i)=k\} = k\,\delta\,h + o(h)\,.$$

If $\delta < \beta$ the expected number of cells in the clone grows exponentially,

$$\mathrm{E}(C(t-i)) = \exp\left((\beta-\delta)\;(t-i)\right)\,,$$

for $t \geq i$. When $\beta = \delta$, the colony will disappear after a finite time $t-i$, whereas for $\beta > \delta$ the colony survives to age $t=\infty$ with probability $(\beta-\delta)/\beta$. A surviving colony satisfies $E(C(t-i)|\,C(t-i)>0) = (\beta/(\beta-\delta))\,\exp\left((\beta-\delta)\,(t-i)\right)$. This process has been analyzed in detail, for example, in Kendall (1948).

2.2.1. Promotion within a clonal expansion

In this section we consider a model for the development of initiated cells based on a time-continuous branching process as in Section 4.2 of Kimmel and Axelrod (2002). In our context, the division rate for initiated cells is higher than for normal cells and their model has to be adapted accordingly. During each division of an initial cell, there is a small chance r_A that a promoted cell is created. Physiologically, promotion turns a pre-neoplastic initiated cell into a neoplastic cell. As before, consider an initiated cell created at time i. During the time interval $[i, i+h]$ the following four possibilities exist

$$\begin{cases} \text{1. a promoted cell and} \\ \text{an initiated cell are created} & \text{, with probability } r_A\beta h + o(h)\,; \\ \text{2. two initiated cells are created} & \text{, with probability } (1-r_A)\beta h + o(h)\,; \\ \text{3. the initiated cell dies and} \\ \text{the expansion stops} & \text{, with probability } \delta h + o(h)\,; \\ \text{4. no division occurs} & \text{, with probability } 1-(\beta+\delta)h + o(h)\,. \end{cases} \tag{3}$$

Let $S_P(t-i)$ be the probability that no neoplastic cell has been created by this process up to time $t \geq i$. Deriving the survival function S_P rather than the more commonly used probability generating function is sufficient for our purposes and simplifies the mathematicas.

Theorem 1. *Suppose the cells created in a homogeneous birth-and-death process with birth rate β and death rate $\delta < \beta$ can acquire a genetic change, A, with probability r_A. Starting from a single cell at time $x = 0$, the probability that none of the descendant cells up to age $x \geq 0$ has changed then satisfies*

$$S_P(x) = \frac{\frac{\rho_2}{(1-r_A)\beta}\,[\rho_1 - (1-r_A)\beta]\left(1-e^{-\Delta x}\right) + [\rho_1 - \rho_2]\; e^{-\Delta x}}{[\rho_1 - (1-r_A)\beta]\left(1-e^{-\Delta x}\right) + [\rho_1 - \rho_2]\; e^{-\Delta x}}\,, \qquad (4)$$

where $\Delta = \sqrt{(\beta-\delta)^2 + 4r_A\beta\delta}$, $\rho_1 = (\beta+\delta+\Delta)/2$ *and* $\rho_2 = (\beta+\delta-\Delta)/2$.

Proof. Partitioning the interval $[0, x]$ into the subintervals $[0, h)$ and $[h, x]$, it follows from (3) that

$$\begin{aligned} S_P(x) &= r_A\beta h \times 0 + (1-r_A)\beta h \times S_P^2(x-h) \\ &\quad +\delta h \times 1 + (1-(\beta+\delta)h) \times S_P(x-h) + o(h) \\ &= (1-r_A)\beta h S_P^2(x-h) + \delta h + (1-(\beta+\delta)h)S_P(x-h) + o(h)\,. \end{aligned}$$

This in turn implies that

$$S_P'(x) = (1-r_A)\beta S_P^2(x) - (\beta+\delta)S_P(x) + \delta\,.$$

This Riccati equation can be solved by putting $S_P(x) = -w'(x)/(w(x)(1-r_A)\beta)$ or $S_P'(x) = -w''(x)/(w(x)(1-r_A)\beta) + (w'(x))^2/(w(x)^2(1-r_A)\beta)$. Our differential equation, after simplification, then becomes

$$-w''(x)/(w(x)(1-r_A)\beta) = \delta + (\beta+\delta)w'(x)/(w(x)(1-r_A)\beta)\,,$$

or

$$w'' + (\beta+\delta)w' + (1-r_A)\beta\delta w = 0\,.$$

The solution $w(x)$ of the above second order linear differential equation with constant coefficients is a linear combination $B_1\exp(-\rho_1 x) + B_2\exp(-\rho_2 x)$, where ρ_1 and ρ_2 are the roots of the characteristic polynomial $\rho^2 + (\beta+\delta)\rho + (1-r_A)\beta\delta$. From $S_P(x) = -w'(x)/(w(x)\,(1-r_A)\beta)$ we now find the general solution (4) to our original differential equation, valid for $x \geq 0$. Using the boundary condition $S_P(0) = 1$ then leads to the result. □

The function S_P behaves like an ordinary survival function in that it starts at $S_P(0) = 1$ and decreases monotonically. However, $S_P(x)$ does not converge to 0 as $x \to \infty$, but rather to $\rho_2/((1-r_A)\beta) = \delta/\beta + o(r_A)$. This is due to the fact that a clone may never gives rise to a neoplastic cell because it dies out. If r_A is small, as is the case in our applications, $\rho_1 = \beta + r_A\beta\delta/(\beta-\delta) + o(r_A)$, $\rho_2 = \delta - r_A\beta\delta/(\beta-\delta) + o(r_A)$, so that $\rho_1 - (1-r_A)\beta = O(r_A)$ is small. This shows that the multipliers of $(1-e^{-\Delta x})$ in (4) are small numbers. As long as $e^{-\Delta x}$ is moderately large, the function $S_P(x)$ remains thus close to 1. A crude approximation of S_P is

$$S_P(t) \approx \frac{\delta r_A(1-e^{-(\beta-\delta)t}) + (\beta-\delta)e^{-(\beta-\delta)t}}{\beta r_A(1-e^{-(\beta-\delta)t}) + (\beta-\delta)e^{-(\beta-\delta)t}}\,,$$

which around age $\log\left[1 + (\beta-\delta)/(r_A\beta)\right]/(\beta-\delta)$ reaches the halfway point between its maximal value of 1 and its minimal value.

2.2.2. Promotion if two genetic changes are required

Suppose two genetic changes, B occurring with probability r_B at each cell division and A with probability r_A, are necessary during the promotion stage and suppose the B-change must precede the A-change. As above, we introduce the survival function for the occurrence of a cell carrying the A-change within the clone of an initiated cell,

$$S_P(t-i) = P\,(\text{no A-cell up to time } t \geq i \,|\, \text{an I-cell has been created at time } i)\ .$$

As an auxiliary function we also need

$$V(t-b) = P\,(\text{no A-cell up to time } t \geq b \,|\, \text{a B-cell has been created at time } b)\ .$$

By an argument analogous to the one given above, we then find

$$\begin{aligned} S_P(x) &= [(1-r_B)\beta h + o(h)]\ S_P^2(x-h) + [r_B\beta h + o(h)]\ V(x-h)S_P(x-h) \\ &\quad + [\delta h + o(h)] + [1-(\beta+\delta)h + o(h)]\ S_P(x-h) \\ V(x) &= [(1-r_A)\beta h + o(h)]\ V^2(x-h) + [\delta h + o(h)] \\ &\quad + [1-(\beta+\delta)h + o(h)]\ V(x-h)\,. \end{aligned}$$

The two functions S_P and V thus satisfy a coupled system of differential equations, namely

$$\begin{aligned} S_P'(x) &= [(1-r_B)\beta]\, S_P^2(x) + r_B\beta V(x)S_P(x) - [\beta+\delta]\, S_P(x) + \delta \\ V'(x) &= [(1-r_A)\beta]\, V^2(x) - [\beta+\delta]\, V(x) + \delta\,. \end{aligned}$$

In Theorem 1 we have found the solution to the equation in $V(x)$, which we can thus eliminate from the equation for $S_P(x)$ by substitution. The resulting Riccati equation in $S_P(x)$ can, however, not be solved in closed form and we would need to use numerical methods. This model can be generalized by incorporating different birth rates for A-cells and B-cells and/or different death rates.

2.3. Survival and age-dependent risk in two-stage carcinogenesis

The two functions $\lambda_I(i)$ and $S_P(t-i)$ describe the rate of creation of pre-neoplastic cells and the probability that no neoplastic cell has been created in a growing clone associated with a pre-neoplastic cell created at age i. We now have to put these two elements together to find the survival function

$$S(t) = P(\text{up to age } t \text{ no neoplastic cell has come into existence})\,.$$

Theorem 2. *Let a cell population be subject to a process creating initiated cells by a non-homogeneous Poisson process with rate function $\lambda_I(t)$. Furthermore, assume that each initiated cell gives rise to a clonal expansion (see Theorem 1) within which $S_P(x)$ gives the survival probability and such that different clonal expansions act independently of each other. The age-dependent risk of acquiring neoplastic cells then satisfies*

$$pobs(t) = \int_0^t \lambda_I(i)\,(-S_P'(t-i))\,di\,. \tag{5}$$

Proof. We consider small intervals of time between $(k-1)t/K$ and kt/K for $1 \le k \le K$. The chance that a new initiated cell is created during this interval is $\lambda_I(kt/K)\,t/K + o(1/K)$. The probability that this new cell does not give rise to a neoplastic cell up to age t is $S_P[(K-k)\,t/K]$. Creation of initiated cells in disjoint intervals is independent and each initiated cell gives rise to a clone that acts independently of other clones. The probability $S(t)$ is thus equal to the product

$$\prod_{k=1}^{K}\left(\left(\lambda_I\left(\frac{kt}{K}\right)\,\frac{t}{K}+o\left(\frac{1}{K}\right)\right)\,S_P\left[\frac{(K-k)t}{K}\right]+\left(1-\lambda_I\left(\frac{kt}{K}\right)\,\frac{t}{K}+o\left(\frac{1}{K}\right)\right)\right),$$

where we take in each interval the possibilities that a new pre-neoplastic cell is born or that no new pre-neoplastic cell is born into account. Rewriting this as

$$S(t)=\exp\left(\sum_{k=1}^{K}\log\left(1-\left(\lambda_I\left(\frac{kt}{K}\right)\frac{t}{K}+o\left(\frac{1}{K}\right)\right)\left(1-S_P\left[\frac{(K-k)t}{K}\right]\right)\right)\right)$$

shows that as $K \to \infty$

$$S(t)=\exp\left(-\int_0^t \lambda_I(i)\,(1-S_P(t-i))\,di\right).$$

To derive this result, note that $\log(1-h) = -h + o(h)$ and interpret the sum as a Riemann integral. The corresponding hazard rate is the one given in (5). □

2.4. Discussion

The physiological parameters needed to determine the incidence rate (5) are the number of cells $N(t)$, mutation probabilities per cell division $r_i, r_j, \ldots$, the number τ of cell divisions per year of normal cells, the probability of the promotion events per cell division $r_A, \ldots$ and the growth characteristics of pre-neoplastic cells, namely δ/β, which is equal to the probability that the clonal expansion resulting from an pre-neoplastic cell dies out and $\beta - \delta$, which describes the exponential growth rate of the clone.

We illustrate (5) under the simplified assumption that the number of cells in the organ is constant and equal to N_0. In that case, the initiation rate is given by (2). With $n = 1$ one then finds

$$\text{pobs}(t) = \tau\, r_1\, N_0\,(1 - S_P(t))\,,$$

that is pobs(t) is close to zero for moderate ages t, then rises to a maximal risk equal to $\tau\, r_1\, N_0\,(1-\delta/\beta)$ and subsequently stays constant. This function pobs(t) describes carcinogenesis as a process that has very small incidence rates at young ages. At some point and relatively suddenly it switches to a steady-state with constant incidence rate.

With two mutations required for initiation ($n = 2$), we have

$$\begin{aligned}
\text{pobs}(t) &= \int_0^t \lambda_I(i)\,(-S_P'(t-i))\,di \\
&= \int_0^t 2\,\tau^2\,(r_1\,r_2)\,N_0\,i\,(-S_P'(t-i))\,di \\
&= 2\,\tau^2\,(r_1\,r_2)\,N_0\,i\,S_P(t-i)\Big|_0^t - \int_0^t 2\,\tau^2\,(r_1\,r_2)\,N_0\,S_P(t-i)\,di \\
&= 2\,\tau^2\,(r_1\,r_2)\,N_0\,t - 2\,\tau^2\,(r_1\,r_2)\,N_0 \int_0^t S_P(i)\,di \\
&= \frac{2\,\tau^2\,(r_1\,r_2)\,N_0}{(1-r_A)\beta}\left(t\,((1-r_A)\beta-\rho_2)\right. \\
&\quad \left. + \log\left(\frac{\rho_1-(1-r_A)\beta}{\Delta} + \frac{(1-r_A)\beta-\rho_2}{\Delta}\,e^{-\Delta t}\right)\right),
\end{aligned}$$

where the constants used are the ones derived in Theorem 1. We remind the reader that $\rho_1 - (1-r_A)\beta = O(r_A)$ is a small probability in actual applications. The first term inside the logarithm has thus a negligible influence as long as $e^{-\Delta t}$ is relatively big. Expanding the logarithm shows that in this case its value is roughly equal to $-((1-r_A)\beta-\rho_2)\,t + \frac{\rho_1-(1-r_A)\beta}{(1-r_A)\beta-\rho_2}\,e^{\Delta t}$. At high ages, the term $e^{-\Delta t}$ becomes negligible and the value of the logarithmic part will be almost constant. Thus, with two initiating mutations the resulting incidence rate is close to zero for young ages, then rises exponentially, proportional to $e^{\Delta t}$, and finally reaches a steady-state with a rate that increases linearly with age.

In more complex cases, (5) can easily be approximated by numerical methods, for example, using the trapezoidal rule

$$\text{pobs}(t) \approx \frac{t}{K}\left(\frac{f(0)}{2} + f\left(\frac{1}{K}\right) + \cdots + f\left(1-\frac{1}{K}\right) + \frac{f(1)}{2}\right),$$

where $f(u) = \lambda_I(tu)\,(-S_P'(t(1-u)))$.

3. The fraction at risk

Models are useful in formulating and refining hypotheses about the process of carcinogenesis, but only if the underlying parameters have either a physiological meaning or describe variation between individuals. Furthermore, the values such parameters can take must be in agreement with current knowledge. Making use of models that take into account such physiological parameters can give hints about features lacking in current models. An example is given by the observation that the mortality due to most forms of cancer shows a peak, typically between the ages of eighty and ninety, and then starts to decline steeply (see for example Herrero-Jimenez et al., 1998). This phenomenon has also been observed in mice that are allowed to live to their full natural lifetime, as reported by Pompei et al. (2001).

The cancer incidence models we discussed above, based on an initiating and a promoting event, have to be adapted in order to provide a model for the turning over of incidence rates at high ages. To achieve this, we present in this section the notion of "a fraction of the population at risk." We propose that a part of the population is genetically, by behavior or simply by a random selection process protected from developing a particular type of cancer. A small proportion of protected individuals at birth can at high age become the majority of the surviving population, thus resulting in almost zero mortality due to that cancer at very high age. In other words, the drop-off occurring at high ages is due to the shrinking of the risk set with aging. About the biological basis of this risk protection one can merely speculate. The fraction at risk model is a binary (at risk/protected) approximation to the statistical distribution of the genetic and environmental fitness of individuals.

The age-specific risk of an incidence would be modified from pobs(t) (see 5) to

$$\text{pobs}(t) \times \frac{\text{Survivors among susceptibles up to age } t}{\text{Survivors in the whole population up to age } t} .$$

The "fraction at risk" hypothesis is vague, but plausible. There are several possible mechanisms creating such an effect. There may, for example, exist parts of the populations with genetic and environmental characteristics that protect these individuals, either by preventing initiation or promotion. In the case of promotion, it would for example suffice if in the protected population the pre-neoplastic cells have a low or zero growth rate. More generally, genetic variation in the population may lead to a variation in the parameters of the two-stage carcinogenesis between different individuals. In this view, the fraction at risk is a simple approximation obtained by using a binary distribution.

In order to apply this concept, we need an expression for the modifier of pobs, that is the relative size of the surviving susceptibles. A reasonable approach consists in partitioning the causes of death into three sets, namely (i) the type of cancer of immediate interest, (ii) causes related via shared risk factors, and (iii) independent causes. The corresponding cause-specific incidence functions describe competing risks and are combined by summation to obtain an incidence rate for all diseases

$$\text{inc}(t) = \text{pobs}(t) + \text{inc}_{\text{rel}}(t) + \text{inc}_{\text{ind}}(t) .$$

If we assume that independent causes act equally on the whole population, whereas related causes only act on the group of susceptibles, the relative size of the susceptibles is equal to

$$\frac{F \exp\left(-\int_0^t \{\text{pobs}(u) + \text{inc}_{\text{rel}}(u) + \text{inc}_{\text{ind}}(u)\}\, du\right)}{F \exp\left(-\int_0^t \{\text{pobs}(u) + \text{inc}_{\text{rel}}(u) + \text{inc}_{\text{ind}}(u)\}\, du\right) + (1-F) \exp\left(-\int_0^t \text{inc}_{\text{ind}}(u)\, du\right)} ,$$

where F denotes the **fraction at risk** of the population at birth. The cause-specific hazard $\text{inc}_{\text{ind}}(t)$ cancels, because of our assumption that it acts in a homogeneous manner in the whole population. The risk due to the cancer and due to related

causes, however, is seen to lead to a decline of the relative size of the susceptibles. We can rewrite this equation as

$$\frac{F}{F + (1 - F)\, \exp\left(\int_0^t \{\text{pobs}(u) + \text{inc}_{\text{rel}}(u)\}\, du\right)}. \tag{6}$$

This is still not sufficiently simple to be useful in practice. A further simplifying hypothesis consists in posing

$$\text{pobs}(u) + \text{inc}_{\text{rel}}(u) = \text{pobs}(u)/f\ ,$$

that is in postulating that the related causes have a cause-specific hazard proportional to the cause-specific hazard of the cancer. The parameter f making its appearance in this equation is the (constant) **fraction of deaths due to the cancer** among all deaths due to either cancer or related causes and thus describes the behavior of competing risks. Substitution in Equation (6) then leads to the final expression for the observable incidence rate

$$\text{pobs}(t) \times \frac{F}{F + (1 - F)\, \exp\left((1/f)\, \int_0^t \text{pobs}(u)\, du\right)}. \tag{7}$$

3.1. Example

For any given model we can adjust the underlying parameters in such a way that the model incidence rate (7) resembles the observed incidence rate obs(t) as closely as possible. Many of the finer distinctions between model parameters and between different models can, however, not be settled by fitting alone, because different, moderately complex, models are often able to fit observed mortalities. Also, several combinations of the parameters of a complicated model may lead to almost equivalent fits. Nevertheless, a convincing argument in favor of multistage models is based on their ability to fit observed incidence rates of many cancers up to about age 80.

The fitting of the two-stage stochastic incidence model and the modification using the parameter couple (F, f) has been explored in Herrero-Jimenez et al. (1998), which is a paper on colon cancer and emphasizes the biological aspects of the model. The use of the model appears to give coherent results. More detailed analysis shows, however, that with a single data set obs(t), the model (7) is too flexible in the sense that many different parameter values are able to explain the data. In order to restrain this flexibility, more quantitative information on the carcinogenesis pathway is required. This includes, for example, more precise measurements on mutation rates, or the number of cells at risk. Another use of the model is in comparing different data sets. If we deal with separate data sets involving the same cancer type $\text{obs}_I(t)$, $\text{obs}_{II}(t)$ and so on, then the flexibility in the model can be restrained by putting conditions on the parameters, for example, by requesting that the physiological parameters in the model be the same for all data sets. In the following example, we will study such a case, namely the mortalities due to lung cancer in the population of European American males born in the 1880's, in the

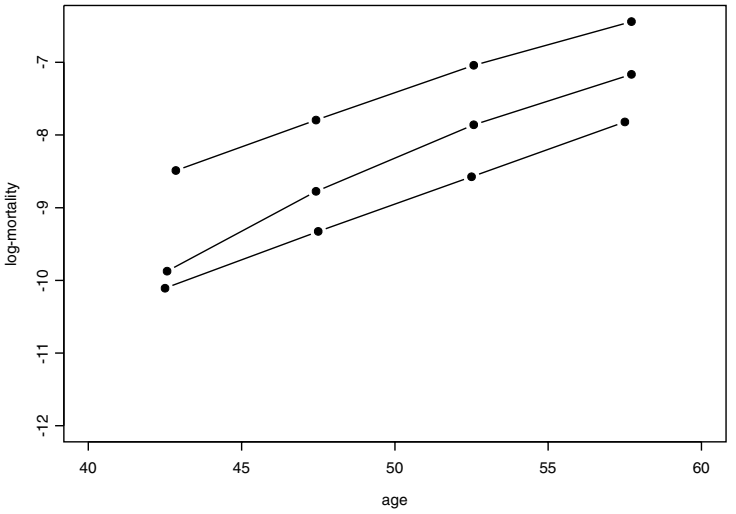

Fig. 1. The three curves are plots of $\log(\mathrm{obs}(t))$ as a function of the age t in the range of 40 to 60 years, which corresponds to the period of initial increase of mortality. The lowest curve is for 1880, the middle one for 1890 and the upper one for 1920. The curves are similar to each other and roughly linear with a slope of 0.15. This indicates that in our model the exponential growth parameter $\beta - \delta$ is equal to 0.15.

1890's and in the 1920's. The data were converted to incidence rates by applying a correction as described in Herrero-Jimenez et al. (1998) for the case of colon cancer. The corresponding observations are summarized in the functions $\mathrm{obs}_{1880}(t)$, $\mathrm{obs}_{1890}(t)$ and $\mathrm{obs}_{1920}(t)$. These data are shown in Figure 2. The large increase in peak-incidence from about 220 per 100000 in the 1880 cohort to about 380 per 100000 in the 1890 cohort to even higher values in the 1920 cohort is self-evident. How can this be explained in terms of our model? If we fit (7) individually to the three observed functions, $\Delta \approx \beta - \delta$ can be found by inspecting the initial rise of the incidences. Figure 1 shows the logarithms of the incidences as a function of age and it indicates a value of about $\beta - \delta = 0.15$. For the purpose of this paper we restrict our modeling efforts to the case of two mutations required for initiation (Eq. 2 with $n = 2$) and one additional genetic change required for promotion (4). Figure 2 shows the results. Good models for the three birth cohorts are obtained when $2\,\tau^2\,(r_1\,r_2)\,N_0 \approx 0.003$ and $r_A \approx 2.5 \times 10^{-6}$. Keeping all these physiological constants the same between the three cohorts we first look for fits with the competing risk parameter f also constant across time. A compromise value is $f = 20\%$. Taking this approach, all parameters except the fraction at risk F are known. Adjusting this last unknown in order to obtain a good agreement between theory and observations, we find $F_{1880} = 37\%$, $F_{1890} = 54\%$ and $F_{1920} \approx 95.6\%$. With time, the fraction at risk increases. Such a marked change over a short period can only be due to environmental factors, most probably smoking habits. The maximal incidence for the cohort born in the 1920's is in this case reached at an age of about 100 years and numbers more than 1350 cases per 100000.

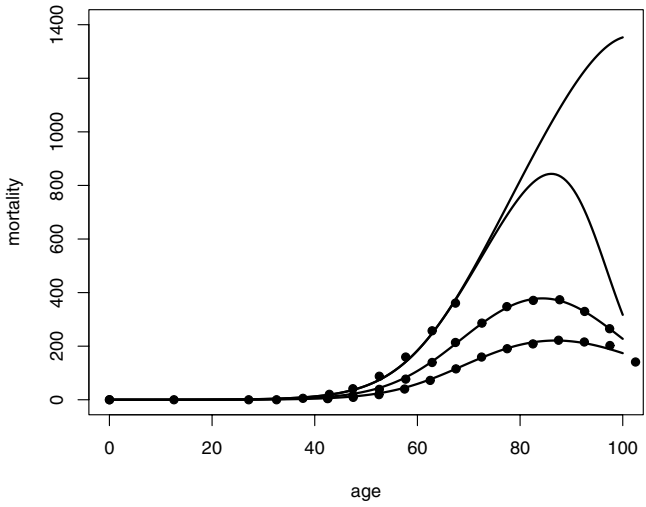

Fig. 2. The incidence rates of lung cancer per 100000 among European American males born in the 1880's (lowest incidence rates), the 1890's (intermediate incidence rates) and the 1920's (highest incidence rates) are shown by the connected points. The model-based incidence rates (7) with $n = 2$ and $m = 1$ are indicated by the superposed curves. When the physiological parameters and the competing risk parameter f are the same in all three cases and only the fraction at risk F varies between cohorts, the cohort born in the 1920's is fitted by the curve with the highest incidence rates. When f, the competing risk parameter is also allowed to change, the 1920's cohort is fitted by the alternative curve with lower incidence rates. In the two other cohorts, the distinction between these two cases is of lesser importance with both resulting in similar curves.

The available data suggests, however, that the competing risk parameter f decreases over time. The incidence curves with $f_{1880} = 26\%$, $f_{1890} = 17\%$ and $f_{1920} \approx 8\%$ and $F_{1880} = 32\%$, $F_{1890} = 58\%$ and $F_{1920} \approx 96.2\%$ given an even closer agreement between theory and observation. In the case of the latter cohort, this combination of population parameters predicts a maximal incidence at age 86 of about 850 cases per 100000. The cohort of those born in the 1920's is particularly difficult to treat since it involves an extrapolation to high ages. The second solution seems preferable also because the overall shape of the incidence curve is similar to the two other curves, whereas in the first instance with fixed f, the age at maximal incidence shifts towards higher ages. The decrease in the competing risk parameter would indicate that even with increased incidences of lung cancer (most probably due to smoking) related diseases also influenced by smoking have grown even more.

4. Conclusions

We have studied a simple modification of a general two-stage carcinogenesis model, which allows one to incorporate variability between individuals into the model. For this purpose, two population parameters, namely the fraction at risk F and the proportion f, had to be introduced. The modified model typically leads to a incidence

rate that turns over at high ages, because the cancer is removing individuals from the risk set. This is an attractive feature, because cancer registries do indeed show this turn over. Further work is needed in the refinement of this model. Other ideas for modeling genetic variation are also very natural and could be tried.

Acknowledgements. The research reported in this paper was supported in part by a grant from the Swiss National Science Foundation. The authors would like to thank the referee for his/her comments, which led to an improved manuscript and pointed out an aspect of the model we had overlooked.

References

Armitage, P., Doll, R.: The age distribution of cancer and a multistage theory of carcinogenesis. Brit. J. Cancer **8**, 1–12 (1954)

Coldman, A.J., Goldie, J.H.: A model for the resistance of tumor cells to cancer chemotherapeutic agents. Math. Biosci. **65**, 291–307 (1983)

Herrero-Jimenez, P., others: Mutation cell kinetics and subpopulations at risk for colon cancer in the united states. Mutation Res. **400**, 553–578 (1998)

Kendall, D.G.: On the generalized "birth-and-death" process. Ann. Math. Statist. **19**, 1–15 (1948)

Kimmel, M., Axelrod, E.A.: Branching Processes in Biology. Springer, 2002

Moolgavkar, S.H., Venzon, D.J.: Two event model for carcinogenesis: Incidence curves for childhood and adult cancer. Math. Biosci. **47**, 55–77 (1979)

Moolgavkar, S.H., Knudson, A.G.: Mutation and Cancer: A model for human carcinogenesis. J. Nat. Cancer Inst. **66**, 1037–1052 (1981)

Moolgavkar, S.H.: Scientific Issues in Quantitative Cancer Risk Assessment. Birkhäuser, Basel Switzerland, 1990

Nordling, N.O.: A new theory on the cancer-inducing mechanism. Brit. J. Cancer **7**, 68–72 (1953)

Pompei, F., Polkanov, M., Wilson, R.: Age distribution of cancer in mice: the incidence turnover at old age. Toxicol. Indust. Health **17**, 7–16 (2001)

Tan, W.Y.: Stochastic Models of Carcinogenesis. Marcel Dekker, New York, USA, 1991

Todorovic, P.: An Introduction to Stochastic Processes and Their Applications. Springer, New York, USA, 1992